Performance Bounds for Expander-Based Compressed Sensing in Poisson Noise

Maxim Raginsky, *Member, IEEE*, Sina Jafarpour, Zachary T. Harmany, *Student Member, IEEE*, Roummel F. Marcia, *Member, IEEE*, Rebecca M. Willett, *Member, IEEE*, and Robert Calderbank, *Fellow, IEEE*

***Abstract*—This paper provides performance bounds for compressed sensing in the presence of Poisson noise using expander graphs. The Poisson noise model is appropriate for a variety of applications, including low-light imaging and digital streaming, where the signal-independent and/or bounded noise models used in the compressed sensing literature are no longer applicable. In this paper, we develop a novel sensing paradigm based on expander graphs and propose a MAP algorithm for recovering sparse or compressible signals from Poisson observations. The geometry of the expander graphs and the positivity of the corresponding sensing matrices play a crucial role in establishing the bounds on the signal reconstruction error of the proposed algorithm. We support our results with experimental demonstrations of reconstructing average packet arrival rates and instantaneous packet counts at a router in a communication network, where the arrivals of packets in each flow follow a Poisson process.**

***Index Terms*—compressive measurement, expander graphs, RIP-1, photon-limited imaging, packet counters**

I. Introduction

The goal of *compressive sampling* or *compressed sensing* (CS) [1], [2] is to replace conventional sampling by a more efficient data acquisition framework, which generally requires fewer sensing resources. This paradigm is particularly enticing whenever the measurement process is costly or constrained in some sense. For example, in the context of photon-limited applications (such as low-light imaging), the photomultiplier tubes used within sensor arrays are physically large and expensive. Similarly, when measuring network traffic flows, the high-speed memory used in packet counters is cost-prohibitive. These problems appear ripe for the application of CS.

The work of M. Raginsky, Z.T. Harmany, and R.M. Willett was supported by NSF CAREER Award No. CCF-06-43947, DARPA Grant No. HR0011-07-1-003, and NSF Grant DMS-08-11062. The work of R. Calderbank and S. Jafarpour was supported in part by NSF under grant DMS 0701226, by ONR under grant N00173-06-1-G006, and by AFOSR under grant FA9550-05-1-0443. The work of R.F. Marcia was supported by NSF Grant No. DMS-08-11062.

M. Raginsky is with the Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708 USA (e-mail: m.raginsky@duke.edu).

S. Jafarpour is with the Department of Computer Science, Princeton University, Princeton, NJ 08540 USA (e-mail: sina@cs.princeton.edu).

Z.T. Harmany and R.M. Willett are with the Department of Electrical and Computer Engineering, Duke University, Durham, NC, 27708, USA (e-mail: zth@duke.edu, willett@duke.edu).

R. Marcia is with the School of Natural Sciences, University of California, Merced, CA 95343 USA (e-mail: rmarcia@ucmerced.edu).

R. Calderbank is with the Department of Computer Science, Duke University, Durham, NC 27708 USA (e-mail: robert.calderbank@duke.edu).

However, photon-limited measurements [3] and arrivals/departures of packets at a router [4] are commonly modeled with a Poisson probability distribution, posing significant theoretical and practical challenges in the context of CS. One of the key challenges is the fact that the measurement error variance scales with the true intensity of each measurement, so that we cannot assume constant noise variance across the collection of measurements. Furthermore, the measurements, the underlying true intensities, and the system models are all subject to certain physical constraints, which play a significant role in performance.

Recent works [5]–[8] explore methods for CS reconstruction in the presence of impulsive, sparse or exponential-family noise, but do not account for the physical constraints associated with a typical Poisson setup and do not contain the related performance bounds emphasized in this paper. In previous work [9], [10], we showed that a Poisson noise model combined with conventional dense CS sensing matrices (properly scaled) yielded performance bounds that were somewhat sobering relative to bounds typically found in the literature. In particular, we found that if the number of photons (or packets) available to sense were held constant, and if the number of measurements, m, was above some critical threshold, then larger m in general led to larger bounds on the error between the true and the estimated signals. This can intuitively be understood as resulting from the fact that dense CS measurements in the Poisson case cannot be zero-mean, and the DC offset used to ensure physical feasibility adversely impacts the noise variance.

The approach considered in this paper hinges, like most CS methods, on reconstructing a signal from compressive measurements by optimizing a sparsity-regularized goodness-of-fit objective function. In contrast to many CS approaches, however, we measure the fit of an estimate to the data using the Poisson log-likelihood instead of a squared error term. This paper demonstrates that the bounds developed in previous work can be improved for some sparsity models by considering alternatives to dense sensing matrices with random entries. In particular, we show that *deterministic* sensing matrices given by scaled adjacency matrices of expander graphs have important theoretical characteristics (especially an ℓ_1 version of the *restricted isometry property* [11]) that are ideally suited to controlling the performance of Poisson CS.

Formally, suppose we have a signal $\theta^* \in \mathbb{R}_+^n$ with known ℓ_1 norm $\|\theta^*\|_1$ (or a known upper bound on $\|\theta^*\|_1$). We aim to find a matrix $\Phi \in \mathbb{R}_+^{m\times n}$ with m, the number

of measurements, as small as possible, so that θ^* can be recovered efficiently from the measured vector $y \in \mathbb{R}_+^m$, which is related to $\Phi\theta^*$ through a Poisson observation model. The restriction that elements of Φ be nonnegative reflects the physical limitations of many sensing systems of interest (e.g., packet routers and counters or linear optical systems). The original approach employed dense random matrices [11], [12]. It has been shown that if the matrix Φ acts nearly isometrically on the set of all k-sparse signals, thus obeying what is now referred to as the Restricted Isometry Property with respect to ℓ_2 norm (RIP-2) [11], then the recovery of θ^* from $\Phi\theta^*$ is indeed possible. It has been also shown that dense random matrices constructed from Gaussian, Bernoulli, or partial Fourier ensembles satisfy the required RIP-2 property with high probability [11].

Adjacency matrices of expander graphs [13] have been recently proposed as an alternative to dense random matrices within the compressed sensing framework, leading to computationally efficient recovery algorithms [14]–[16]. It has been shown that variations of the standard recovery approaches such as *basis pursuit* [2] and *matching pursuit* [17] are consistent with the expander sensing approach and can recover the original sparse signal successfully [18], [19]. In the presence of Gaussian or sparse noise, random dense sensing and expander sensing are known to provide similar performance in terms of the number of measurements and recovery computation time. Berinde et al. proved that expander graphs with sufficiently large expansion are near-isometries on the set of all k-sparse signals in the ℓ_1 norm; this is referred as a Restricted Isometry Property for ℓ_1 norm (RIP-1) [18]. Furthermore, expander sensing requires less storage whenever the signal is sparse in the canonical basis, while random dense sensing provides slightly tighter recovery bounds [16].

The approach described in this paper consists of the following key elements:

- expander sensing matrices and the RIP-1 associated with them;
- a reconstruction objective function which explicitly incorporates the Poisson likelihood;
- a countable collection of candidate estimators; and
- a penalty function defined over the collection of candidates, which satisfies the Kraft inequality and which can be used to promote sparse reconstructions.

In general, the penalty function is selected to be small for signals of interest, which leads to theoretical guarantees that errors are small with high probability for such signals. In this paper, exploiting the RIP-1 property and the non-negativity of the expander-based sensing matrices, we show that, in contrast to random dense sensing, expander sensing empowered with a maximum *a posteriori* (MAP) algorithm can approximately recover the original signal in the presence of Poisson noise, and we prove bounds which quantify the MAP performance. As a result, in the presence of Poisson noise, expander graphs not only provide general storage advantages, but they also allow for efficient MAP recovery methods with performance guarantees comparable to the best k-term approximation of the original signal. Finally, the bounds are tighter than those for specific dense matrices proposed by Willett and Raginsky [9], [10] whenever the signal is sparse in the canonical domain, in that a log term in the bounds in [10] is absent from the bounds presented in this paper.

A. Relationship with dense sensing matrices for Poisson CS

In recent work, the authors established performance bounds for CS in the presence of Poisson noise using dense sensing matrices based on appropriately shifted and scaled Rademacher ensembles [9], [10]. Several features distinguish that work from the present paper:

- The dense sensing matrices used in [9], [10] require more memory to store and more computational resources to apply to a signal in a reconstruction algorithm. The expander-based approach described in this paper, in contrast, is more efficient.
- The expander-based approach described in this paper works *only* when the signal of interest is sparse in the canonical basis. In contrast, the dense sensing matrices used in [9], [10] can be applied to arbitrary sparsity bases (though the proof technique there needs to be altered slightly to accommodate sparsity in the canonical basis).
- The bounds in *both* this paper and [9], [10] reflect a sobering tradeoff between performance and the number of measurements collected. In particular, more measurements (after some critical minimum number) can actually *degrade* performance as a limited number of events (e.g., photons) are distributed among a growing number of detectors, impairing the SNR of the measurements.

B. Notation

Nonnegative reals (respectively, integers) will be denoted by $\mathbb{R}_+$ (respectively, $\mathbb{Z}_+$). Given a vector $u \in \mathbb{R}^n$ and a set $S \subseteq \{1, \ldots, n\}$, we will denote by u^S the vector obtained by setting to zero all coordinates of u that are in S^c, the complement of S: $\forall 1 \leq i \leq n, u_i^S = u_i 1_{\{i \in S\}}$. Given some $1 \leq k \leq n$, let S be the set of positions of the k largest (in magnitude) coordinates of u. Then $u^{(k)} \triangleq u^S$ will denote the *best k-term approximation* of u (in the canonical basis of $\mathbb{R}^n$), and

$$\sigma_k(u) \triangleq \|u - u^{(k)}\|_1 = \sum_{i \in S^c} |u_i|$$

will denote the resulting ℓ_1 approximation error. The ℓ_0 quasinorm measures the number of nonzero coordinates of u: $\|u\|_0 \triangleq \sum_{i=1}^n 1_{\{u_i \neq 0\}}$. For a subset $S \subseteq \{1, \ldots, n\}$ we will denote by I_S the vector with components $1_{\{i \in S\}}$, $1 \leq i \leq n$. Given a vector u, we will denote by u^+ the vector obtained by setting to zero all negative components of u: for all $1 \leq i \leq n$, $u_i^+ = \max\{0, u_i\}$. Given two vectors $u, v \in \mathbb{R}^n$, we will write $u \succeq v$ if $u_i \geq v_i$ for all $1 \leq i \leq n$. If $u \succeq \alpha I_{\{1,\ldots,n\}}$ for some $\alpha \in \mathbb{R}$, we will simply write $u \succeq \alpha$. We will write $\succ$ instead of $\succeq$ if the inequalities are strict for all i.

C. Organization of the paper

This paper is organized as follows. In Section II, we summarize the existing literature on expander graphs applied

to compressed sensing and the RIP-1 property. Section III describes how the problem of compressed sensing with Poisson noise can be formulated in a way that explicitly accounts for nonnegativity constraints and flux preservation (i.e., we cannot detect more events than have occurred); this section also contains our main theoretical result bounding the error of a sparsity penalized likelihood reconstruction of a signal from compressive Poisson measurements. These results are illustrated and further analyzed in Section IV, in which we focus on the specific application of efficiently estimating packet arrival rates. Several technical discussions and proofs have been relegated to the appendices.

II. Background on expander graphs

We start by defining an *unbalanced bipartite vertex-expander graph*.

Definition II.1. We say that a bipartite simple graph $G = (A, B, E)$ with (regular) left degree[1] d is a (k, ϵ)-*expander* if, for any $S \subset A$ with $|S| \leq k$, the set of neighbors $\mathcal{N}(S)$ of S has size $|\mathcal{N}(S)| > (1-\epsilon)d\,|S|$.

Figure 1 illustrates such a graph. Intuitively a bipartite graph is an expander if any sufficiently small subset of its variable nodes has a sufficiently large neighborhood. In the CS setting, A (resp., B) will correspond to the components of the original signal (resp., its compressed representation). Hence, for a given $|A|$, a "high-quality" expander should have $|B|$, d, and ϵ as small as possible, while k should be as close as possible to $|B|$. The following proposition, proved using the probabilistic method [20], is well-known in the literature on expanders:

Proposition II.2 (Existence of high-quality expanders)**.** For any $1 \leq k \leq \frac{n}{2}$ and any $\epsilon \in (0,1)$, there exists a (k,ϵ)-expander with left degree $d = O\left(\frac{\log(n/k)}{\epsilon}\right)$ and right set size $m = O\left(\frac{k\log(n/k)}{\epsilon^2}\right)$.

Unfortunately, there is no explicit construction of expanders from Definition II.1. However, it can be shown that, with high probability, any d-regular random graph with

$$d = O\left(\frac{\log(n/k)}{\epsilon}\right) \text{ and } m = O\left(\frac{k\log(n/k)}{\epsilon^2}\right)$$

satisfies the required expansion property. Moreover, the graph may be assumed to be *right-regular* as well, i.e., every node in B will have the same (right) degree D [21]. Counting the number of edges in two ways, we conclude that

$$|E| = |A|d = |B|D \qquad \Longrightarrow \qquad D = O\left(\frac{n}{k}\right).$$

Thus, in practice it may suffice to use random bipartite regular graphs instead of expanders[2]. Moreover, there exists an explicit construction for a class of expander graphs that comes very close to the guarantees of Proposition II.2. This construction, due to Guruswami et al. [23], uses Parvaresh-Vardy codes [24] and has the following guarantees:

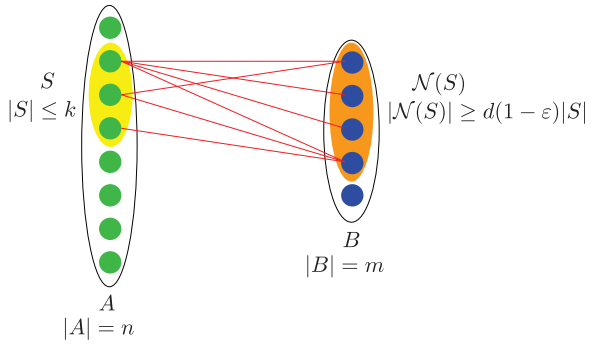


Fig. 1. A (k,ϵ)-expander. In this example, the green nodes correspond to A, the blue nodes correspond to B, the yellow oval corresponds to the set $S \subset A$, and the orange oval corresponds to the set $\mathcal{N}(S) \subset B$. There are three colliding edges.

Proposition II.3 (Explicit construction of high-quality expanders)**.** For any positive constant β, and any n, k, ϵ, there exists a deterministic explicit construction of a (k,ϵ)-expander graph with $d = O\left(\left(\frac{\log n}{\epsilon}\right)^{\frac{1+\beta}{\beta}}\right)$ and $m = O(d^2k^{1+\beta})$.

Expanders have been recently proposed as a means of constructing efficient compressed sensing algorithms [15], [18], [19], [22]. In particular, it has been shown that any n-dimensional vector that is k-sparse can be fully recovered using $O\left(k\log\left(\frac{n}{k}\right)\right)$ measurements in $O\left(n\log\left(\frac{n}{k}\right)\right)$ time [15], [19]. It has been also shown that, even in the presence of noise in the measurements, if the noise vector has low ℓ_1 norm, expander-based algorithms can approximately recover any k-sparse signal [16], [18], [19]. One reason why expander graphs are good sensing candidates is that the adjacency matrix of any (k,ϵ)-expander almost preserves the ℓ_1 norm of any k-sparse vector [18]. In other words, if the adjacency matrix of an expander is used for measurement, then the ℓ_1 distance between two sufficiently sparse signals is preserved by measurement. This property is known as the "Restricted Isometry Property for ℓ_1 norms" or the "RIP-1" property. Berinde et al. have shown that this condition is sufficient for sparse recovery using ℓ_1 minimization [18].

The precise statement of the RIP-1 property, whose proof can be found in [15], goes as follows:

Lemma II.4 (RIP-1 property of the expander graphs)**.** Let F be the $m \times n$ adjacency matrix of a (k,ϵ) expander graph G. Then for any k-sparse vector $x \in \mathbb{R}^n$ we have:

$$(1-2\epsilon)d\|x\|_1 \leq \|Fx\|_1 \leq d\|x\|_1 \tag{1}$$

The following proposition is a direct consequence of the above RIP-1 property. It states that if, for any almost k-sparse vector[3] u, there exists a vector v whose ℓ_1 norm is close to that of u, and if Fv approximates Fu, then v also approximates u. Our results of Section III exploit the fact

[1] That is, each node in A has the same number of neighbors in B.

[2] Briefly, we can first generate a random left-regular graph with left degree d (by choosing each edge independently). That graph is, with overwhelming probability, an expander graph. Then, given an expander graph which is only left-regular, a paper by Guruswami et al. [22] shows how to construct an expander graph with almost the same parameters, which is both left-regular and right-regular.

[3] By "almost sparsity" we mean that the vector has at most k significant entries.

that the proposed MAP decoding algorithm outputs a vector satisfying the two conditions above, and hence approximately recovers the desired signal.

Proposition II.5. Let F be the adjacency matrix of a $(2k, \epsilon)$-expander and u, v be two vectors in $\mathbb{R}^n$, such that

$$\|u\|_1 \geq \|v\|_1 - \Delta$$

for some $\Delta > 0$. Then $\|u - v\|_1$ is upper-bounded by

$$\|u - v\|_1 \leq \frac{1-2\epsilon}{1-6\epsilon}\left(2\sigma_k(u) + \Delta\right) + \frac{2}{d(1-6\epsilon)}\|Fu - Fv\|_1.$$

In particular, if we let $\epsilon = 1/16$, then we get the bound

$$\|u - v\|_1 \leq 4\sigma_k(u) + \frac{4}{d}\|Fu - Fv\|_1 + 2\Delta.$$

Proof: See Appendix B. ∎

For future convenience, we will introduce the following piece of notation. Given n and $1 \leq k \leq n/4$, we will denote by $G_{k,n}$ a $(2k, 1/16)$-expander with left set size n whose existence is guaranteed by Proposition II.2. Then $G_{k,n} = (A, B, E)$ has

$$|A| = n, \quad |B| = m = O(k \log(n/k)), \quad d = O(\log(n/k)).$$

III. Compressed sensing in the presence of Poisson Noise

A. Problem statement

We wish to recover an unknown vector $\theta^* \in \mathbb{R}_+^n$ of Poisson intensities from a measured vector $y \in \mathbb{Z}_+^m$, sensed according to the Poisson model

$$y \sim \text{Poisson}(\Phi\theta^*), \tag{2}$$

where $\Phi \in \mathbb{R}_+^{m \times n}$ is a positivity-preserving sensing matrix[4]. That is, for each $j \in \{1, \ldots, m\}$, y_j is sampled independently from a Poisson distribution with mean $(\Phi\theta^*)_j$:

$$\mathbb{P}_{\Phi\theta^*}(y) = \prod_{j=1}^{m} \mathbb{P}_{(\Phi\theta^*)_j}(y_j), \tag{3}$$

where, for any $z \in \mathbb{Z}_+$ and $\lambda \in \mathbb{R}_+$, we have

$$\mathbb{P}_\lambda(z) \triangleq \begin{cases} \dfrac{\lambda^z}{z!}e^{-\lambda} & \text{if } \lambda > 0 \\ 1_{\{z=0\}} & \text{otherwise} \end{cases}, \tag{4}$$

where the $\lambda = 0$ case is a consequence of the fact that

$$\lim_{\lambda \to 0} \frac{\lambda^z}{z!}e^{-\lambda} = 1_{\{z=0\}}.$$

We assume that the ℓ_1 norm of θ^* is known, $\|\theta^*\|_1 = L$ (although later we will show that this assumption can be relaxed). We are interested in designing a sensing matrix Φ and an estimator $\hat{\theta} = \hat{\theta}(y)$, such that θ^* can be recovered with small expected ℓ_1 risk

$$R\left(\hat{\theta}, \theta^*\right) \triangleq \mathbb{E}_{\Phi\theta^*}\|\hat{\theta} - \theta^*\|_1,$$

where the expectation is taken w.r.t. the distribution $\mathbb{P}_{\Phi\theta^*}$.

[4]Our choice of this observation model as opposed to a "shot-noise" model based on Φ operating on Poisson observations of θ^* is discussed in Appendix A.

B. The proposed estimator and its performance

To recover θ^*, we will use a penalized Maximum Likelihood Estimation (pMLE) approach. Let us choose a convenient $1 \leq k \leq n/4$ and take Φ to be the normalized adjacency matrix of the expander $G_{k,n}$ (cf. Section II for definitions): $\Phi \triangleq F/d$. Moreover, let us choose a finite or countable set Θ_L of candidate estimators $\theta \in \mathbb{R}_+^n$ with $\|\theta\|_1 \leq L$, and a *penalty* $\text{pen} : \Theta_L \to \mathbb{R}_+$ satisfying the *Kraft inequality*[5]

$$\sum_{\theta \in \Theta_L} e^{-\text{pen}(\theta)} \leq 1. \tag{5}$$

For instance, we can impose less penalty on sparser signals or construct a penalty based on any other prior knowledge about the underlying signal.

With these definitions, we consider the following *penalized maximum likelihood estimator (pMLE)*:

$$\hat{\theta} \triangleq \operatorname*{argmin}_{\theta \in \Theta_L} \left[-\log \mathbb{P}_{\Phi\theta}(y) + 2\,\text{pen}(\theta)\right] \tag{6}$$

One way to think about the procedure in (6) is as a Maximum *a posteriori* Probability (MAP) algorithm over the set of estimates Θ_L, where the likelihood is computed according to the Poisson model (4) and the penalty function corresponds to a negative log prior on the candidate estimators in Θ_L.

Our main bound on the performance of the pMLE is as follows:

Theorem III.1. Let Φ be the normalized adjacency matrix of $G_{k,n}$, let $\theta^* \in \mathbb{R}_+^n$ be the original signal compressively sampled in the presence of Poisson noise, and let $\hat{\theta}$ be obtained through (6). Then

$$R\left(\hat{\theta}, \theta^*\right) \leq 4\sigma_k(\theta^*) + 8\sqrt{L \min_{\theta \in \Theta_L}\left[\text{KL}(\mathbb{P}_{\Phi\theta^*} \parallel \mathbb{P}_{\Phi\theta}) + 2\,\text{pen}(\theta)\right]}, \tag{7}$$

where

$$\text{KL}(\mathbb{P}_g \| \mathbb{P}_h) \triangleq \sum_{y \in \mathbb{Z}_+^m} \mathbb{P}_g(y) \log \frac{\mathbb{P}_g(y)}{\mathbb{P}_h(y)}$$

is the Kullback–Leibler divergence (relative entropy) between $\mathbb{P}_g$ and $\mathbb{P}_h$ [25].

Proof: Since $\hat{\theta} \in \Theta_L$, we have $L = \|\theta^*\|_1 \geq \|\hat{\theta}\|_1$. Hence, using Proposition II.5 with $\Delta = 0$, we can write

$$\|\theta^* - \hat{\theta}\|_1 \leq 4\sigma_k(\theta^*) + 4\|\Phi(\theta^* - \hat{\theta})\|_1.$$

Taking expectations, we obtain

$$\begin{aligned} R\left(\hat{\theta}, \theta^*\right) &\leq 4\sigma_k(\theta^*) + 4\mathbb{E}_{\Phi\theta^*}\|\Phi(\theta^* - \hat{\theta})\|_1 \\ &\leq 4\sigma_k(\theta^*) + 4\sqrt{\mathbb{E}_{\Phi\theta^*}\|\Phi(\theta^* - \hat{\theta})\|_1^2} \end{aligned} \tag{8}$$

where the second step uses Jensen's inequality. Using Lemmas C.1 and C.2 in Appendix C, we have

$$\mathbb{E}_{\Phi\theta^*}\|\Phi(\theta^* - \hat{\theta})\|_1^2 \leq 4L \min_{\theta \in \Theta_L}\left[\text{KL}(\mathbb{P}_{\Phi\theta^*} \parallel \mathbb{P}_{\Phi\theta}) + 2\,\text{pen}(\theta)\right]$$

[5]Many penalization functions can be modified slightly (e.g., scaled appropriately) to satisfy the Kraft inequality. All that is required is a finite collection of estimators (i.e., Θ_L) and an associated prefix code for each candidate estimate in Θ_L. For instance, this would certainly be possible for a total variation penalty, though the details are beyond the scope of this paper.

Substituting this into (8), we obtain (7). ∎

The bound of Theorem III.1 is an *oracle inequality*: it states that the ℓ_1 error of $\hat{\theta}$ is (up to multiplicative constants) the sum of the k-term approximation error of θ^* plus $\sqrt{L}$ times the minimum penalized relative entropy error over the set of candidate estimators Θ_L. The first term in (7) is smaller for sparser θ^*, and the second term is smaller when there is a $\theta \in \Theta_L$ which is simultaneously a good approximation to θ^* (in the sense that the distributions $\mathbb{P}_{\Phi\theta^*}$ and $\mathbb{P}_{\Phi\theta}$ are close) and has a low penalty.

Remark III.2. So far we have assumed that the ℓ_1 norm of θ^* is known *a priori*. If this is not the case, we can still estimate it with high accuracy using noisy compressive measurements. Observe that, since each measurement y_j is a Poisson random variable with mean $(\Phi\theta^*)_j$, $\sum_j y_j$ is Poisson with mean $\|\Phi\theta^*\|_1$. Therefore, $\sqrt{\sum_j y_j}$ is approximately normally distributed with mean $\approx \sqrt{\|\Phi\theta^*\|_1}$ and variance $\approx \frac{1}{4}$ [26, Sec. 6.2].[6] Hence, Mill's inequality [27, Thm. 4.7] guarantees that, for every positive t,

$$\Pr\left[\left|\sqrt{\sum_j y_j} - \sqrt{\|\Phi\theta^*\|_1}\right| > t\right] \lesssim \frac{e^{-2t^2}}{\sqrt{2\pi}t},$$

where $\lesssim$ is meant to indicate the fact that this is only an approximate bound, with the approximation error controlled by the rate of convergence in the central limit theorem. Now we can use the RIP-1 property of the expander graphs obtain the estimates

$$\left(\sqrt{\sum_j y_j} - t\right)^2 \le \|\Phi\theta^*\|_1 \le \|\theta^*\|_1,$$

and

$$\frac{\left(\sqrt{\sum_j y_j} + t\right)^2}{(1-2\epsilon)} \ge \frac{\|\Phi\theta^*\|_1}{(1-2\epsilon)} \ge \|\theta^*\|_1$$

that hold with (approximate) probability at least $1 - (\sqrt{2\pi}t)^{-1}e^{-2t^2}$.

C. A bound in terms of ℓ_1 error

The bound of Theorem III.1 is not always useful since it bounds the ℓ_1 risk of the pMLE in terms of the relative entropy. A bound purely in terms of ℓ_1 errors would be more desirable. However, it is not easy to obtain without imposing extra conditions either on θ^* or on the candidate estimators in Θ_L. This follows from the fact that the divergence $\mathrm{KL}(\mathbb{P}_{\Phi\theta^*}\|\mathbb{P}_{\Phi\theta})$ may take the value $+\infty$ if there exists some y such that $\mathbb{P}_{\Phi\theta}(y) = 0$ but $\mathbb{P}_{\Phi\theta^*}(y) > 0$.

One way to eliminate this problem is to impose an additional requirement on the candidate estimators in Θ_L: There exists some $c > 0$, such that

$$\Phi\theta \succeq c, \qquad \forall\theta \in \Theta_L \tag{9}$$

Under this condition, we will now develop a risk bound for the pMLE purely in terms of the ℓ_1 error.

[6]This observation underlies the use of variance-stabilizing transforms.

Theorem III.3. Suppose that all the conditions of Theorem III.1 are satisfied. In addition, suppose that the set Θ_L satisfies the condition (9). Then

$$R\left(\hat{\theta}, \theta^*\right) \le 4\sigma_k(\theta^*) + 8\sqrt{L \min_{\theta\in\Theta_L}\left[\frac{\|\theta^* - \theta\|_1^2}{c} + 2\,\mathrm{pen}(\theta)\right]}. \tag{10}$$

Proof: Using Lemma C.3 in Appendix C, we get the bound

$$\mathrm{KL}(\mathbb{P}_{\Phi\theta^*}\|\mathbb{P}_{\Phi\theta}) \le \frac{1}{c}\|\theta^* - \theta\|_1^2, \qquad \forall\theta \in \Theta_L.$$

Substituting this into Eq. (7), we get (10). ∎

Remark III.4. Because every $\theta \in \Theta_L$ satisfies $\|\theta\|_1 \le L$, the constant c cannot be too large. In particular, if (9) holds, then for every $\theta \in \Theta_L$ we must have

$$\|\Phi\theta\|_1 \ge m \min_j (\Phi\theta)_j \ge mc.$$

On the other hand, by the RIP-1 property we have $\|\Phi\theta\|_1 \le \|\theta\|_1 \le L$. Thus, a necessary condition for (9) to hold is $c \le L/m$. Since $m = O(k\log(n/k))$, the best risk we may hope to achieve under some condition like (9) is on the order of

$$R\left(\hat{\theta}, \theta^*\right) \le 4\sigma_k(\theta^*) + C\sqrt{\min_{\theta\in\Theta_L}\left[k\log(n/k)\|\theta - \theta^*\|_1^2 + L\,\mathrm{pen}(\theta)\right]} \tag{11}$$

for some constant C, e.g., by choosing $c \propto \frac{L}{k\log(n/k)}$. Effectively, this means that, under the positivity condition (9), the ℓ_1 error of $\hat{\theta}$ is the sum of the k-term approximation error of θ^* plus $\sqrt{m} = \sqrt{k\log(n/k)}$ times the best penalized ℓ_1 approximation error. The first term in (11) is smaller for sparser θ^*, and the second term is smaller when there is a $\theta \in \Theta_L$ which is simultaneously a good ℓ_1 approximation to θ^* and has a low penalty.

D. Empirical performance

Here we present a simulation study that validates our method. In this experiment, compressive Poisson observations are collected of a randomly generated sparse signal passed through the sensing matrix generated from an adjacency matrix of an expander. We then reconstruct the signal by utilizing an algorithm that minimizes the objective function in (6), and assess the accuracy of this estimate. We repeat this procedure over several trials to estimate the average performance of the method.

More specifically, we generate our length-n sparse signal θ^* through a two-step procedure. First we select k elements of $\{1, \ldots, n\}$ uniformly at random, then we assign these elements an intensity I. All other components of the signal are set to zero. For these experiments, we chose a length $n = 100{,}000$ and varied the sparsity k among three different choices of 100, 500, and 1,000 for two intensity levels I of 10,000 and 100,000. We then vary the number m of Poisson observations from 100 to 20,000 using an expander graph sensing matrix with degree $d = 8$. Recall that the sensing matrix is normalized

such that the total signal intensity is divided amongst the measurements, hence the seemingly high choices of I.

To reconstruct the signal, we utilize the SPIRAL-ℓ_1 algorithm [28] which solves (6) when $\mathrm{pen}(\theta) = \tau\|\theta\|_1$. We design the algorithm to optimize over the continuous domain $\mathbb{R}_+^n$ instead of the discrete set Θ_L. This is equivalent to the proposed pMLE formulation in the limit as the discrete set of estimates becomes increasingly dense in the set of all $\theta \in \mathbb{R}_+^n$ with $\|\theta\|_1 \leq L$, i.e., we quantize this set on an ever finer scale, increasing the bit allotment to represent each θ. In this high-resolution limit, the Kraft inequality requirement (5) on the penalty $\mathrm{pen}(\theta)$ will translate to $\int e^{-\mathrm{pen}(\theta)} d\theta < \infty$. If we select a penalty proportional to the negative log of a prior probability distribution for θ, this requirement will be satisfied. From a Bayesian perspective, the ℓ_1 penalty arises by assuming each component θ_i is drawn i.i.d. from a zero-mean Laplace prior $p(\theta_i) = e^{-|\theta_i|/b}/2b$. Hence the regularization parameter τ is inversely related to the scale parameter b of the prior, as a larger τ (smaller b) will promote solutions with more zero-valued components.

This relaxation results in a computationally tractable convex program over a continuous domain, albeit implemented on a machine with finite precision. The SPIRAL algorithm utilizes a sequence of quadratic subproblems derived by using a second-order Taylor expansion of the Poisson log-likelihood at each iteration. These subproblems are made easier to solve by using a separable approximation whereby the second-order Hessian matrix is approximated by a scaled identity matrix. For the particular case of the ℓ_1 penalty, these subproblems can be solved quickly, exactly, and noniteratively by a soft-thresholding rule.

After reconstruction, we assess the estimate $\hat{\theta}$ according to the normalized ℓ_1 error $\|\theta^* - \hat{\theta}\|_1 / \|\theta^*\|_1$. We select the regularization weighting τ in the SPIRAL-ℓ_1 algorithm to minimize this quantity for each randomly generated experiment indexed by (I, k, m). To assure that the results are not biased in our favor by only considering a single random experiment for each (I, k, m), we repeat this experiment several times. The averaged reconstruction accuracy over 10 trials is presented in Figure 2.

These results show that the proposed method is able to accurately estimate sparse signals when the signal intensity is sufficiently high; however, the performance of the method degrades for lower signal strengths. More interesting is the behavior as we vary the number of measurements. There is a clear phase transition where accurate signal reconstruction becomes possible, however the performance gently degrades with the number of measurements since there is a lower signal-to-noise ratio per measurement. This effect is more pronounced at lower intensity levels, as we more quickly enter the regime where only a few photons are collected per measurement. These findings support the error bounds developed in Section III-B.

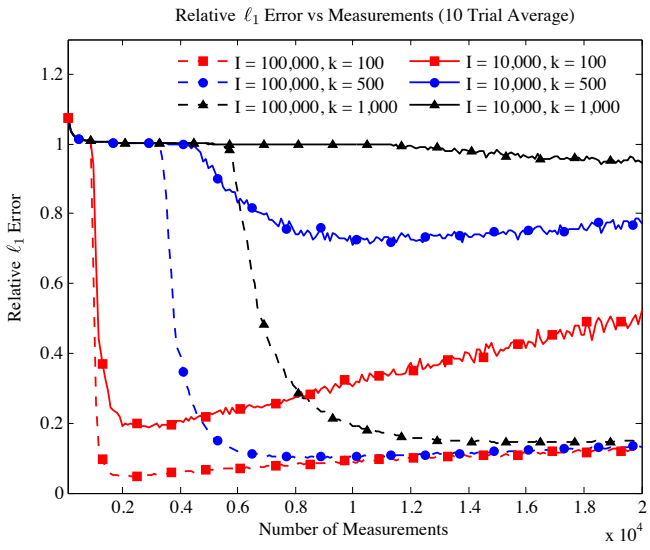


Fig. 2. Average performance (as measured by the normalized ℓ_1 error $\|\theta^* - \hat{\theta}\|_1 / \|\theta^*\|_1$) for the proposed expander-based observation method for recovering sparse signals under Poisson noise. In this experiment, we sweep over a range of measurements and consider a few sparsity (k) and intensity (I) levels of the true signal.

IV. Application: Estimating Packet Arrival Rates

This section describes an application of the pMLE estimator of Section III: an indirect approach for reconstructing average packet arrival rates and instantaneous packet counts for a given number of streams (or flows) at a router in a communication network, where the arrivals of packets in each flow are assumed to follow a Poisson process. All packet counting must be done in hardware at the router, and any hardware implementation must strike a delicate balance between speed, accuracy, and cost. For instance, one could keep a dedicated counter for each flow, but, depending on the type of memory used, one could end up with an implementation that is either fast but expensive and unable to keep track of a large number of flows (e.g., using SRAMs, which have low access times, but are expensive and physically large) or cheap and high-density but slow (e.g., using DRAMs, which are cheap and small, but have longer access times) [29], [30].

However, there is empirical evidence [31], [32] that flow sizes in IP networks follow a *power-law* pattern: just a few flows (say, 10%) carry most of the traffic (say, 90%). Based on this observation, several investigators have proposed methodologies for estimating flows using a small number of counters by either (a) keeping track only of the flows whose sizes exceed a given fraction of the total bandwidth (the approach suggestively termed "focusing on the elephants, ignoring the mice") [29] or (b) using sparse random graphs to aggregate the raw packet counts and recovering flow sizes using a message passing decoder [30].

We consider an alternative to these approaches based on Poisson CS, assuming that the underlying Poisson rate vector is sparse or approximately sparse — and, in fact, it is the approximate sparsity of the rate vector that mathematically describes the power-law behavior of the average packet counts. The goal is to maintain a compressed summary of the process sample paths using a small number of counters, such that it

is possible to reconstruct both the total number of packets in each flow and the underlying rate vector. Since we are dealing here with Poisson streams, we would like to push the metaphor further and say that we are "focusing on the whales, ignoring the minnows."

A. Problem formulation

We wish to monitor a large number n of packet flows using a much smaller number m of counters. Each flow is a homogeneous Poisson process (cf. [4] for details pertaining to Poisson processes and networking applications). Specifically, let $\lambda^* \in \mathbb{R}_+^n$ denote the vector of rates, and let U denote the random process $U = \{U_t\}_{t\in\mathbb{R}_+}$ with sample paths in $\mathbb{Z}_+^n$, where, for each $i \in \{1,\ldots,n\}$, the ith component of U is a homogeneous Poisson process with the rate of λ_i arrivals per unit time, and all the component processes are mutually conditionally independent given λ.

The goal is to estimate the unknown rate vector λ based on y. We will focus on performance bounds for power-law network traffic, i.e., for λ^* belonging to the class

$$\Sigma_{\alpha,L_0} \triangleq \left\{\lambda \in \mathbb{R}_+^n : \|\lambda\|_1 = L_0; \sigma_k(\lambda) = O(k^{-\alpha})\right\} \quad (12)$$

for some $L_0 > 0$ and $\alpha \geq 1$, where the constant hidden in the $O(\cdot)$ notation may depend on L_0. Here, α is the power-law exponent that controls the tail behavior; in particular, the extreme regime $\alpha \to +\infty$ describes the fully sparse setting. As in Section III, we assume the total arrival rate $\|\lambda^*\|_1$ to be known (and equal to a given L_0) in advance, but this assumption can be easily dispensed with (cf. Remark III.2).

As before, we evaluate each candidate estimator $\widehat{\lambda} = \widehat{\lambda}(y)$ based on its expected ℓ_1 risk,

$$R\left(\widehat{\lambda}, \lambda^*\right) = \mathbb{E}_{\lambda^*}\|\widehat{\lambda} - \lambda^*\|_1.$$

B. Two estimation strategies

We consider two estimation strategies. In both cases, we let our measurement matrix F be the adjacency matrix of the expander $G_{k,n}$ for a fixed $k \leq n/4$ (see Section II for definitions). The first strategy, which we call the *direct method*, uses standard expander-based CS to construct an estimate of λ^*. The second is the pMLE strategy, which relies on the machinery presented in Section III and can be used when only the rates are of interest.

1) The direct method: In this method, which will be used as a "baseline" for assessing the performance of the pMLE, the counters are updated in discrete time, every τ time units. Let $x = \{x_\nu\}_{\nu\in\mathbb{Z}_+}$ denote the sampled version of U, where $x_\nu \triangleq U_{\nu\tau}$. The update takes place as follows. We have a binary matrix $F \in \{0,1\}^{m\times n}$, and at each time ν let $y_\nu = Fx_\nu$. In other words, y is obtained by passing a sampled n-dimensional homogeneous Poisson process with rate vector λ through a linear transformation F.

The direct method uses expander-based CS to obtain an estimate $\widehat{x}_\nu$ of x_ν from $y_\nu = Fx_\nu$, followed by letting

$$\widehat{\lambda}_\nu^{\mathrm{dir}} = \frac{\widehat{x}_\nu^+}{\nu\tau}. \quad (13)$$

This strategy is based on the observation that $x_\nu/(\nu\tau)$ is the maximum-likelihood estimator of λ^*. To obtain $\widehat{x}_\nu$, we need to solve the convex program

$$\text{minimize } \|u\|_1 \qquad \text{subject to } Fu = y_\nu$$

which can be cast as a linear program [33]. The resulting solution $\widehat{x}_\nu$ may have negative coordinates,[7] hence the use of the $(\cdot)^+$ operation in (13). We then have the following result:

Theorem IV.1.

$$R\left(\widehat{\lambda}_\nu^{\mathrm{dir}}, \lambda^*\right) \leq 4\sigma_k(\lambda^*) + \frac{\|(\lambda^*)^{1/2}\|_1}{\sqrt{\nu\tau}}, \quad (14)$$

where $(\lambda^*)^{1/2}$ is the vector with components $\sqrt{\lambda_i^*}, \forall i$.

Remark IV.2. Note that the error term in (14) is $O(1/\sqrt{\nu})$, assuming everything else is kept constant, which coincides with the optimal rate of the ℓ_1 error decay in parametric estimation problems.

Proof: We first observe that, by construction, $\widehat{x}_\nu$ satisfies the relations $F\widehat{x}_\nu = Fx_\nu$ and $\|\widehat{x}_\nu\|_1 \leq \|x_\nu\|_1$. Hence,

$$\begin{aligned}\mathbb{E}\|\widehat{x}_\nu - \nu\tau\lambda^*\|_1 &\leq \mathbb{E}\|\widehat{x}_\nu - x_\nu\|_1 + \mathbb{E}\|x_\nu - \nu\tau\lambda^*\|_1 \\ &\leq 4\mathbb{E}\sigma_k(x_\nu) + \mathbb{E}\|x_\nu - \nu\tau\lambda^*\|_1 \quad (15)\end{aligned}$$

where the first step uses the triangle inequality, while the second step uses Proposition II.5 with $\Delta = 0$. To bound the first term in (15), let $S \subset \{1,\ldots,n\}$ denote the positions of the k largest entries of λ^*. Then, by definition of the best k-term representation,

$$\sigma_k(x_\nu) \leq \|x_\nu - x_\nu^S\|_1 = \sum_{i\in S^c}|x_{\nu,i}| = \sum_{i\in S^c} x_{\nu,i}.$$

Therefore,

$$\mathbb{E}\sigma_k(x_\nu) \leq \mathbb{E}\left[\sum_{i\in S^c} x_{\nu,i}\right] = \nu\tau\sum_{i\in S^c}\lambda_i^* \equiv \nu\tau\sigma_k(\lambda^*).$$

To bound the second term, we can use concavity of the square root, as well as the fact that each $x_{\nu,i} \sim \mathrm{Poisson}(\nu\tau\lambda_i^*)$, to write

$$\begin{aligned}\mathbb{E}\|x_\nu - \nu\tau\lambda^*\|_1 &= \mathbb{E}\left[\sum_{i=1}^N |x_{n,i} - \nu\tau\lambda_i^*|\right] \\ &= \mathbb{E}\left[\sum_{i=1}^N \sqrt{(x_{n,i} - \nu\tau\lambda_i^*)^2}\right] \\ &\leq \sum_{i=1}^n \sqrt{\mathbb{E}(x_{\nu,i} - \nu\tau\lambda_i^*)^2} = \sum_{i=1}^n \sqrt{\nu\tau\lambda_i^*}.\end{aligned}$$

Now, it is not hard to show that $\|\widehat{x}_\nu^+ - \nu\tau\lambda^*\|_1 \leq \|\widehat{x}_\nu - \nu\tau\lambda^*\|_1$. Therefore,

$$R\left(\widehat{\lambda}_\nu^{\mathrm{dir}}, \lambda^*\right) \leq \frac{\mathbb{E}\|\widehat{x}_\nu - \nu\tau\lambda^*\|_1}{\nu\tau} \leq 4\sigma_k(\lambda^*) + \frac{\|(\lambda^*)^{1/2}\|_1}{\sqrt{\nu\tau}},$$

which proves the theorem. ∎

[7] Khajehnejad et al. [34] have recently proposed the use of perturbed adjacency matrices of expanders to recover nonnegative sparse signals.

2) The penalized MLE approach: In the penalized MLE approach the counters are updated in a slightly different manner. Here the counters are still updated in discrete time, every τ time units; however, each counter $i \in \{1, \cdots, m\}$ is updated at times $\left(\nu\tau + \frac{i}{m}\tau\right)_{\nu \in \mathbb{Z}_+}$, and only aggregates the packets that have arrived during the time period $\left[\nu\tau + \frac{i-1}{m}\tau, \nu\tau + \frac{i}{m}\tau\right)$. Therefore, in contrast to the direct method, here each arriving packet is registered by at most one counter. Furthermore, since the packets arrive according to a homogeneous Poisson process, conditioned on the vector λ^*, the values measured by distinct counters are independent[8]. Therefore, the vector of counts at time ν obeys

$$y_\nu \sim \text{Poisson}(\Phi\theta^*) \qquad \text{where} \qquad \theta^* = \frac{\nu\tau d}{m}\lambda^*$$

which is precisely the sensing model we have analyzed in Section III.

Now assume that the total average arrival rate $\|\lambda^*\|_1 = L_0$ is known. Let Λ be a finite or a countable set of candidate estimators with $\|\lambda\|_1 \leq L_0$ for all $\lambda \in \Lambda$, and let $\text{pen}(\cdot)$ be a penalty functional satisfying the Kraft inequality over Λ. Given ν and τ, consider the scaled set

$$\Lambda_{\nu,\tau} \triangleq \frac{\nu\tau d}{m}\Lambda \equiv \left\{\frac{\nu\tau d}{m}\lambda : \lambda \in \Lambda\right\}$$

with the same penalty function, $\text{pen}\left(\frac{\nu\tau d}{m}\lambda\right) = \text{pen}(\lambda)$ for all $\lambda \in \Lambda$. We can now apply the results of Section III. Specifically, let

$$\widehat{\lambda}_\nu^{\text{pMLE}} \triangleq \frac{m\,\hat{\theta}}{\nu\tau d},$$

where $\hat{\theta}$ is the corresponding pMLE estimator obtained according to (6). The following theorem is a consequence of Theorem III.3 and the remark following it:

Theorem IV.3. If the set Λ satisfies the strict positivity condition (9), then there exists some absolute constant $C > 0$, such that

$$R\left(\widehat{\lambda}_\nu^{\text{pMLE}}, \lambda^*\right) \leq 4\sigma_k(\lambda^*) + C\sqrt{\min_{\lambda\in\Lambda}\left[k\log(n/k)\|\lambda - \lambda^*\|_1^2 + \frac{k\,L_0\,\text{pen}(\lambda)}{\nu\tau}\right]}. \tag{16}$$

We now develop risk bounds under the power-law condition. To this end, let us suppose that λ^* is a member of the power-law class $\Sigma_{L_0,\alpha}$ defined in (12). Fix a small positive number δ, such that $L_0/\sqrt{\delta}$ is an integer, and define the set

$$\Lambda \triangleq \left\{\lambda \in \mathbb{R}_+^n : \|\lambda\|_1 \leq L_0; \lambda_i \in \{s\sqrt{\delta}\}_{s=0}^{L_0/\sqrt{\delta}}, \forall i\right\}$$

These will be our candidate estimators of λ^*. We can define the penalty function $\text{pen}(\lambda) \asymp \|\lambda\|_0 \log(\delta^{-1})$. For any $\lambda \in \Sigma_{\alpha,L_0}$ and any $1 \leq r \leq n$ we can find some $\lambda^{(r)} \in \Lambda$, such that $\|\lambda^{(r)}\|_0 \asymp r$ and

$$\|\lambda - \lambda^{(r)}\|_1^2 \asymp r^{-2\alpha} + r\,\delta.$$

[8]The independence follows from the fact that if $X_1, \cdots, X_m$ are conditionally independent random variables, then for any choice of functions $g_1, \cdots, g_m$, the random variables $g_1(X_1), \cdots, g_m(X_m)$ are also conditionally independent.

Here we assume that δ is sufficiently small, so that the penalty term $\frac{k\,r\log(\delta^{-1})}{\nu\tau}$ dominates the quantization error $r\,\delta$. In order to guarantee that the penalty function satisfies Kraft's inequality, we need to ensure that

$$\sum_{r=1}^{n}\sum_{\substack{\lambda^{(r)}\in\Lambda \\ \|\lambda^{(r)}\|_0 = r}} \delta^r \leq 1.$$

For every fixed r, there are exactly $\binom{n}{r}$ subspaces of dimension r, and each subspace contains exactly $\left(\frac{L_0}{\sqrt{\delta}}\right)^r$ distinct elements of Λ. Therefore, as long as

$$\delta \leq (2n\,L_0)^{-2}, \tag{17}$$

then

$$\sum_{r=1}^{n}\binom{n}{r}\left(L_0\sqrt{\delta}\right)^r \leq \sum_{r=0}^{n}\left(n\,L_0\sqrt{\delta}\right)^r \leq \sum_{r=1}^{n}\frac{1}{2^r} \leq 1,$$

and Kraft's inequality is satisfied.

Using the fact that $k\log(n/k) = O(kd)$, we can bound the minimum over $\lambda \in \Lambda$ in (16) from above by

$$\begin{aligned}\min_{1\leq r\leq n}\left[kdr^{-2\alpha} + \frac{r\,k\log(\delta^{-1})}{\nu\tau}\right] &= O\left(k\,d^{\frac{1}{2\alpha+1}}\right)\left(\frac{\log(\delta^{-1})}{\nu\tau}\right)^{\frac{2\alpha}{2\alpha+1}} \\ &= O\left(k\,d^{\frac{1}{2\alpha+1}}\right)\left(\frac{\log n}{\nu\tau}\right)^{\frac{2\alpha}{2\alpha+1}}\end{aligned}$$

We can now particularize Theorem IV.3 to the power-law case:

Theorem IV.4.

$$\sup_{\lambda^*\in\Sigma_{\alpha,L_0}} R\left(\widehat{\lambda}_\nu^{\text{pMLE}}, \lambda^*\right) = O(k^{-\alpha}) + O\left(k^{\frac{1}{2}}\,d^{\frac{1}{4\alpha+2}}\right)\left(\frac{\log n}{\nu\tau}\right)^{\frac{\alpha}{2\alpha+1}},$$

where the constants implicit in the $O(\cdot)$ notation depend on L_0 and α.

Note that the risk bound here is slightly worse than the benchmark bound of Theorem IV.1. However, it should be borne in mind that this bound is based on Theorem III.3, rather than on the potentially much tighter oracle inequality of Theorem III.1, since our goal was to express the risk of the pMLE purely in terms of the ℓ_1 approximation properties of the power-law class Σ_{α,L_0}. In general, we will expect the actual risk of the pMLE to be much lower than what the conservative bound of Theorem IV.4 predicts. Indeed, as we will see in Section IV-D, the pMLE approach obtains higher empirical accuracy than the direct method. But first we show how the pMLE can be approximated efficiently with proper preprocessing of the observed counts y_ν based on the structure of $G_{k,n}$.

C. Efficient pMLE approximation

In this section we present an efficient algorithm for approximating the pMLE estimate. The algorithm consists of two phases: (1) first, we preprocess y_ν to isolate a subset A_1 of $A = \{1, \ldots, n\}$ which is sufficiently small and is guaranteed to contain the locations of the k largest entries of λ^* (the whales); (2) then we construct a set Λ of candidate estimators whose support sets lie in A_1, together with an appropriate penalty, and perform pMLE over this reduced set.

The success of this approach hinges on the assumption that the magnitude of the smallest whale is sufficiently large compared to the magnitude of the largest minnow. Specifically, we make the following assumption: Let $S \subset A$ contain the locations of the k largest coordinates of λ^*. Then we require that

$$\min_{i \in S} \lambda_i^* > 9D \left\| \lambda^* - \lambda^{*(k)} \right\|_\infty. \tag{18}$$

Recall that $D = O\left(\frac{nd}{m}\right) = O\left(\frac{n}{k}\right)$ is the right degree of the expander graph. One way to think about (18) is in terms of a signal-to-noise ratio, which must be strictly larger than $9\,D$. We also require $\nu\tau$ to be sufficiently large, so that

$$\frac{\nu\tau}{m} D \left\| \lambda^* - \lambda^{*(k)} \right\|_\infty \geq \frac{\log(mn)}{2}. \tag{19}$$

Finally, we perturb our expander a bit as follows: choose an integer $k' > 0$ so that

$$k' \geq \max\left\{ \frac{16(kd+1)}{15d}, 2k \right\}. \tag{20}$$

Then we replace our original $(2k, 1/16)$-expander $G_{k,n}$ with left-degree d with a $(k', 1/16)$-expander $G'_{k',n}$ with the same left degree.The resulting procedure, displayed below as Algorithm 1, has the following guarantees:

Algorithm 1 Efficient pMLE approximation algorithm

Input: Measurement vector y_ν, and the sensing matrix F.
Output: An approximation $\widehat{\lambda}$

Let B_1 consist of the locations of the kd largest elements of y_ν and let $B_2 = B \backslash B_1$.
Let A_2 contain the set of all variable nodes that have at least one neighbor in B_2 and let $A_1 = A \backslash A_2$.
Construct a candidate set of estimators Λ with support in A_1 and a penalty $\mathrm{pen}(\cdot)$ over Λ.
Output the pMLE $\widehat{\lambda}$.

Theorem IV.5. Suppose the assumptions (18), (19), and (20) hold. Then with probability at least $1 - \frac{1}{n}$ the set A_1 constructed by Algorithm 1 has the following properties: (1) $S \subset A_1$; (2) $|A_1| \leq kd$; (3) A_1 can be found in time $O(m \log m + nd)$.

Proof: (1) First fix a measurement node $j \in B$. Recall that $y_{\nu,j}$ is a Poisson random variable with mean $\frac{\nu\tau}{m}\left(F\lambda^*\right)_j$. By the same argument as in Remark III.2, $\sqrt{y_{\nu,j}}$ is approximately normally distributed with mean $\approx \sqrt{\frac{\nu\tau}{m}\left(F\lambda^*\right)_j}$, and with variance $\approx \frac{1}{4}$. Hence, it follows from Mill's inequality and the union bound that for every positive t

$$\Pr\left[\exists j : \left| \sqrt{y_{\nu,j}} - \sqrt{\frac{\nu\tau}{m}\left(F\lambda^*\right)_j} \right| > t \right] \lesssim \frac{m e^{-2t^2}}{\sqrt{2\pi} t}.$$

If j is a neighbor of S, then $\left(F\lambda^*\right)_j \geq \min_{i \in S} \lambda_i^*$; whereas if j is not connected to S, then $\left(F\lambda^*\right)_j \leq D \left\| \lambda^* - \lambda^{*(k)} \right\|_\infty$. Hence, by setting $t = \sqrt{\frac{\log(mn)}{2}}$ (where w.l.o.g. we assume that $t \geq 1$), we conclude that, with probability at least $1 - \frac{1}{n}$, for every measurement node j the following holds:

- If j is a neighbor of S, then

$$\sqrt{y_{\nu,j}} \geq \sqrt{\frac{\nu\tau}{m} \min_{i \in S} \lambda_i^*} - \sqrt{\frac{\log(mn)}{2}}.$$

- If j is not connected to S, then

$$\sqrt{y_{\nu,j}} \leq \sqrt{\frac{\nu\tau}{m} D \left\| \lambda^* - \lambda^{*(k)} \right\|_\infty} + \sqrt{\frac{\log(mn)}{2}}.$$

Consequently, by virtue of (18) and (19), with probability at least $1 - \frac{1}{n}$ every element of y_ν that is a neighbor of S has larger magnitude than every element of y_ν that is not a neighbor of S.

(2) Suppose, to the contrary, that $|A_1| > kd$. Let $A_1' \subseteq A_1$ be any subset of size $kd + 1$. Now, Lemma 3.6 in [34] states that, provided $\epsilon \leq 1 - 1/d$, then every (ℓ, ϵ)-expander with left degree d is also a $(\ell(1-\epsilon)d, 1-1/d)$-expander with left degree d. We apply this result to our $(k', 1/16)$-expander, where k' satisfies (20), to see that it is also a $(kd+1, 1-1/d)$-expander. Therefore, for the set A_1' we must have $|\mathcal{N}(A_1')| \geq |A_1'| = kd + 1$. On the other hand, $\mathcal{N}(A_1') \subset B_1$, so $|\mathcal{N}(A_1')| \leq kd$. This is a contradiction, hence we must have $|A_1| \leq kd$.

(3) Finding the sets B_1 and B_2 can be done in $O(m \log m)$ time by sorting y_ν. The set A_1 can then can be found in time $O(nd)$, by sequentially eliminating all nodes connected to each node in B_2. ∎

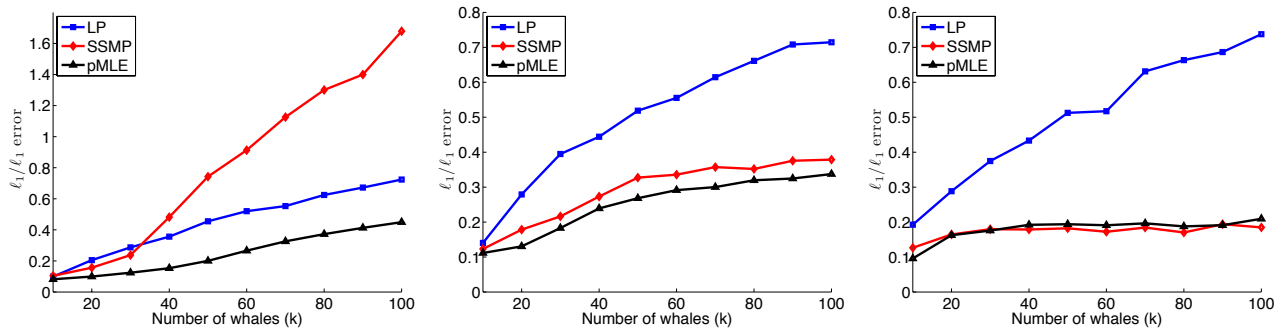
Having identified the set A_1, we can reduce the pMLE optimization only to those candidates whose support sets lie in A_1. More precisely, if we originally start with a sufficiently rich class of estimators $\tilde{\Lambda}$, then the new feasible set can be reduced to

$$\Lambda \triangleq \left\{ \lambda \in \tilde{\Lambda} : \mathrm{Supp}(\lambda) \subset A_1 \right\}.$$

Hence, by extracting the set A_1, we can significantly reduce the complexity of finding the pMLE estimate. If $|\Lambda|$ is small, the optimization can be performed by brute-force search in $O(|\Lambda|)$ time. Otherwise, since $|A_1| \leq kd$, we can use the quantization technique from the preceding section with quantizer resolution $\sqrt{\delta}$ to construct a Λ of size at most $(L_0/\sqrt{\delta})^{kd}$. In this case, we can even assign the uniform penalty

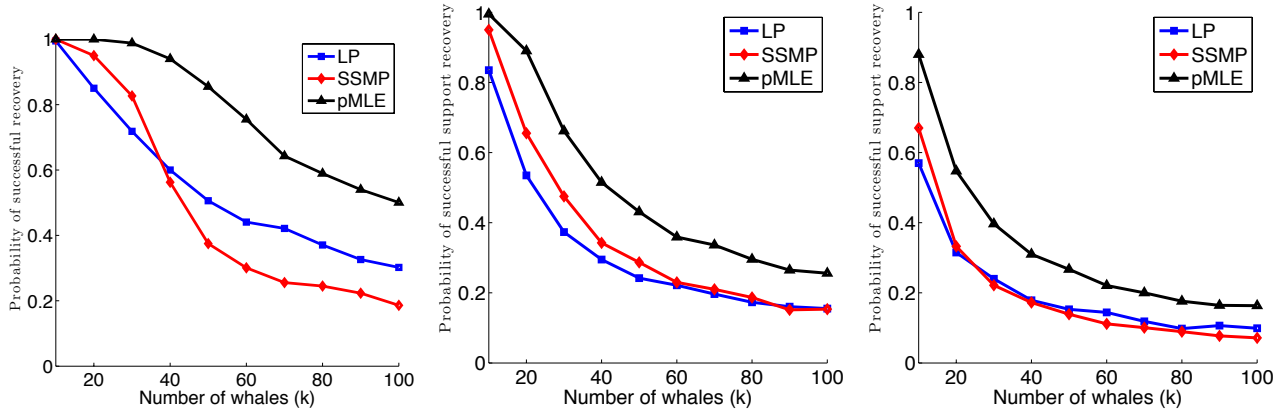
$$\mathrm{pen}(\lambda) = \log|\Lambda| = O\left(k \log(n/k) \log(\delta^{-1})\right),$$

which amounts to a vanilla MLE over Λ.



(a) $\alpha = 1$ (b) $\alpha = 1.5$ (c) $\alpha = 2$

Fig. 3. Relative ℓ_1 error as a function of number of whales k, for ℓ_1-magic (LP), SSMP and pMLE for different choices of the power-law exponent α. The number of flows $n = 5000$, the number of counters $m = 800$, and the number of updates is 40.



(a) $\alpha = 1$ (b) $\alpha = 1.5$ (c) $\alpha = 2$

Fig. 4. Probability of successful support recovery as a function of number of whales k, for ℓ_1-magic (LP), SSMP and pMLE for different choices of the power-law exponent α. The number of flows $n = 5000$, the number of counters $m = 800$, and the number of updates is 40.

D. Empirical performance

Here we compare penalized MLE with ℓ_1-magic [35], a universal ℓ_1 minimization method, and with SSMP [36], an alternative method that employs combinatorial optimization. ℓ_1-magic and SSMP both compute the "direct" estimator. The pMLE estimate is computed using Algorithm 1 above. For the ease of computation, the candidate set Λ is approximated by the convex set of all positive vectors with bounded ℓ_1 norm, and the CVX package [37], [38] is used to directly solve the pMLE objective function with $\mathrm{pen}(\theta) = \|\theta\|_1$.

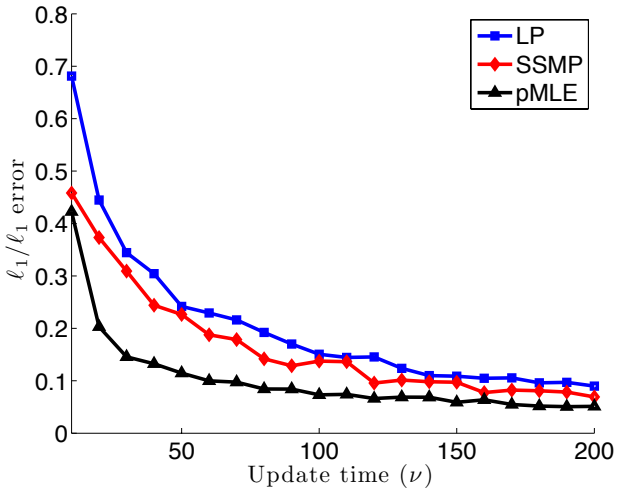
Figures 3(a) through 5(b) report the results of numerical experiments, where the goal is to identify the k largest entries in the rate vector from the measured data. Since a random graph is, with overwhelming probability, an expander graph, each experiment was repeated 30 times using independent sparse random graphs with $d = 8$.

We also used the following process to generate the rate vector. First, given the power-law exponent α, the magnitudes of the k whales were chosen according to a power-law distribution with parameter α. The positions of the k whales were then chosen uniformly at random. Finally the $n - k$ minnows were sampled independently from a $\mathcal{N}(0, 10^{-6})$ distribution (negative samples were replaced by their absolute values). Thus, given the locations of the k whales, their magnitudes decay according to a *truncated* power law (with the cut-off at k), while the magnitudes of the minnows represent a noisy background. Figure 3 shows the relative ℓ_1 error ($\|\lambda - \hat{\lambda}_\nu\|_1/\|\lambda\|_1$) of the three above algorithms as a function of k. Note that in all cases $\alpha = 1$, $\alpha = 1.5$, and $\alpha = 2$, the pMLE algorithm provides lower ℓ_1 errors. Similarly, Figure 4 reports the probability of exact recovery as a function of k. Again, it turns out that in all three cases the pMLE algorithm has higher probability of exact support recovery compared to the two direct algorithms.

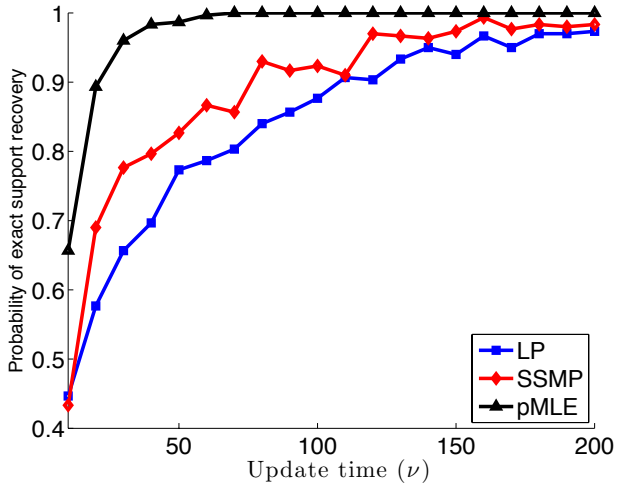
We also analyzed the impact of changing the number of updates on the accuracy of the three above algorithms. The results are demonstrated in Figure 5. Here we fixed the number of whales to $k = 30$, and changed the number of updates from 10 to 200. It turned out that as the number of updates ν increases, the relative ℓ_1 errors of all three algorithms decrease and their probability of exact support recovery consistently increase. Moreover, the pMLE algorithm always outperforms the ℓ_1-magic (LP), and SSMP algorithms.

V. Conclusions

In this paper we investigated expander-based sensing as an alternative to dense random sensing in the presence of Poisson noise. Even though the Poisson model is essential in



(a) Relative ℓ_1 error as a function of number of updates ν.



(b) Probability of successful support recovery as a function of number of updates ν.

Fig. 5. Performance of ℓ_1-magic, SSMP and pMLE algorithms as a function of the number of updates ν. The number of flows $n = 5000$, the number of counters $m = 800$, and the number of whales is $k = 30$. There are k whales whose magnitudes are assigned according to a power-law distribution with $\alpha = 1$, and the remaining entries are minnows with magnitudes determined by a $\mathcal{N}(0, 10^{-6})$ random variable.

some applications, it presents several challenges as the noise is not bounded, or even as concentrated as Gaussian noise, and is signal-dependent. Here we proposed using normalized adjacency matrices of expander graphs as an alternative construction of sensing matrices, and we showed that the binary nature and the RIP-1 property of these matrices yield provable consistency for a MAP reconstruction algorithm.

The compressed sensing algorithms based on Poisson observations and expander-graph sensing matrices provide a useful mechanism for accurately and robustly estimating a collection of flow rates with relatively few counters. These techniques have the potential to significantly reduce the cost of hardware required for flow rate estimation. While previous approaches assumed packet counts matched the flow rates exactly or that flow rates were i.i.d., the approach in this paper accounts for the Poisson nature of packet counts with relatively mild assumptions about the underlying flow rates (i.e., that only a small fraction of them are large).

The "direct" estimation method (in which first the vector of flow counts is estimated using a linear program, and then the underlying flow rates are estimated using Poisson maximum likelihood) is juxtaposed with an "indirect" method (in which the flow rates are estimated in one pass from the compressive Poisson measurements using penalized likelihood estimation).

The methods in this paper, along with related results in this area, are designed for settings in which the flow rates are sufficiently stationary, so that they can be accurately estimated in a fixed time window. Future directions include extending these approaches to a more realistic setting in which the flow rates evolve over time. In this case, the time window over which packets should be counted may be relatively short, but this can be mitigated by exploiting estimates of the flow rates in earlier time windows.

Appendix A
Observation Models in Poisson Inverse Problems

In (2) and all the subsequent analysis in this paper, we assume

$$y \sim \text{Poisson}(\Phi\theta^*).$$

However, one might question how accurately this models the physical systems of interest, such as a photon-limited imaging system or a router. In particular, we may prefer to think of only a small number of events (e.g., photons or packets) being incident upon our system, and the system then rerouting those events to a detector. In this appendix, we compare the statistical properties of these two models. Let $z_{j,i}$ denote the number of events traveling from location i in the source (θ^*) to location j on the detector. Also, in this appendix let us assume Φ is a stochastic matrix, i.e., each column of Φ sums to one; in general, most elements of Φ are going to be less than one. Physically, this assumption means that every event incident on the system hits some element of the detector array. Armed with these assumptions, we can think of $\Phi_{j,i}$ as the probability of events from location i in θ^* being transmitted to location j in the observation vector y.

We consider two observation models:

$$\begin{aligned}
\text{Model A:} \qquad z_{j,i} &\sim \text{Poisson}(\Phi_{j,i}\theta^*_i) \\
y_j &\triangleq \sum_{i=1}^{n} z_{j,i} \\
\text{Model B:} \qquad w &\sim \text{Poisson}(\theta^*) \\
\{z_{j,i}\}_{i=1}^{n} &\sim \text{Multinomial}(w_i, \{\Phi_{j,i}\}_{i=1}^{n}) \\
y_j &\triangleq \sum_{i=1}^{n} z_{j,i},
\end{aligned}$$

where in both models all the components $z_{j,i}$ of z are mutually conditionally independent given the appropriate parameters. Model A roughly corresponds to the model we consider throughout the paper; Model B corresponds to considering Poisson realizations with intensity θ^* (denoted w) incident upon our system and then redirected to different detector

elements via Φ. We model this redirection process with a multinomial distribution. While the model $y \sim \text{Poisson}(\Phi\theta^*)$ is slightly different from Model A, the following analysis will provide valuable insight into discrete event counting systems.

We now show that the distribution of z is the same in Models A and B. First note that

$$y_j \equiv \sum_{i=1}^{n} z_{j,i} \qquad \text{and} \qquad w_i \equiv \sum_{j=1}^{m} z_{j,i}. \tag{21}$$

Under Model A, we have

$$\begin{aligned} p(z|\theta^*) &= \prod_{i=1}^{n}\prod_{j=1}^{m} \frac{e^{-\Phi_{j,i}\theta_i^*}(\Phi_{j,i}\theta_i^*)^{z_{j,i}}}{z_{j,i}!} \\ &= \prod_{i=1}^{n}\left(\prod_{j=1}^{m}\frac{\Phi_{j,i}^{z_{j,i}}}{z_{j,i}!}\right) e^{-\sum_{j=1}^{m}\Phi_{j,i}\theta_i^*}(\theta_i^*)^{\sum_{j=1}^{m} z_{j,i}} \\ &= \prod_{i=1}^{n}\left(\prod_{j=1}^{m}\frac{\Phi_{j,i}^{z_{j,i}}}{z_{j,i}!}\right) e^{-\theta_i^*}(\theta_i^*)^{w_i} \end{aligned} \tag{22}$$

where in the last step we used (21) and the assumption that $\sum_{j=1}^{m}\Phi_{j,i} = 1$. Under Model B, we have

$$\begin{aligned} p(z|w) &= \begin{cases} \displaystyle\prod_{i=1}^{n} w_i! \prod_{j=1}^{m}\frac{\Phi_{j,i}^{z_{j,i}}}{z_{j,i}!}, & \text{if } \displaystyle\sum_{j=1}^{m} z_{j,i} = w_i \,\forall i \\ 0, & \text{otherwise} \end{cases} \\ p(w|\theta^*) &= \prod_{i=1}^{n}\frac{e^{-\theta_i^*}(\theta_i^*)^{w_i}}{w_i!} \\ p(z|\theta^*) &= \sum_{v\in\mathbb{Z}_+^m:\sum_j z_{j,i}=v_i} p(z|v)p(v|\theta^*) \\ &= \prod_{i=1}^{n}\prod_{j=1}^{m} w_i!\frac{\Phi_{j,i}^{z_{j,i}}}{z_{j,i}!}\frac{e^{-\theta_i^*}(\theta_i^*)^{w_i}}{w_i!} \\ &= \prod_{i=1}^{n}\left(\prod_{j=1}^{m}\frac{\Phi_{j,i}^{z_{j,i}}}{z_{j,i}!}\right) e^{-\theta_i^*}(\theta_i^*)^{w_i}. \end{aligned} \tag{23}$$

The fourth line uses (21). Since (22) and (23) are the same, we have shown that Models A and B are statistically equivalent. While Model B may be more intuitively appealing based on our physical understanding of how these systems operate, using Model A for our analysis and algorithm development is just as accurate and mathematically more direct.

Appendix B
Proof of Proposition II.5

Let $y = u - v$, let $S \subset \{1,\dots,n\}$ denote the positions of the k largest (in magnitude) coordinates of y, and enumerate the complementary set S^c as $i_1, i_2, \dots, i_{n-k}$ in decreasing order of magnitude of $|y_{i_j}|, j = 1,\dots,n-k$. Let us partition the set S^c into adjacent blocks $S_1,\dots,S_t$, such that all blocks (but possibly S_t) have size k. Also let $S_0 = S$. Let $\tilde{F}$ be a submatrix of F containing rows from $\mathcal{N}(S)$. Then, following the argument of Berinde et al. [18], which also goes back to Sipser and Spielman [21], we have the following chain of inequalities:

$$\begin{aligned} \|Fy\|_1 &\geq \|\tilde{F}y\|_1 \\ &\geq \|\tilde{F}y_S\|_1 - \sum_{i=1}^{t}\sum_{(j,l)\in E: j\in S_i,\, l\in\mathcal{N}(S)} |y_j| \\ &\geq d(1-2\epsilon)\|y_S\|_1 - \sum_{i=1}^{t}\sum_{(j,l)\in E: j\in S_i,\, l\in\mathcal{N}(S)} \frac{\|y_{S_{i-1}}\|_1}{k} \\ &\geq d(1-2\epsilon)\|y_S\|_1 - 2kd\epsilon\sum_{i=1}^{t}\frac{\|y_{S_{i-1}}\|_1}{k} \\ &\geq d(1-2\epsilon)\|y_S\|_1 - 2d\epsilon\|y\|_1. \end{aligned}$$

Most of the steps are straightforward consequences of the definitions, the triangle inequality, or the RIP-1 property. The fourth inequality follows from the following fact. Since we are dealing with a $(2k,\epsilon)$-expander and since $|S\cup S_i| \leq 2k$ for every $i = 0,\dots,t$, we must have $|\mathcal{N}(S\cup S_i)| \geq d(1-\epsilon)|S\cup S_i|$. Therefore, at most $2kd\epsilon$ edges can cross from each S_i to $\mathcal{N}(S)$. From the above estimate, we obtain

$$\|Fu - Fv\|_1 + 2d\epsilon\|y\|_1 \geq (1-2\epsilon)d\|y_S\|_1. \tag{24}$$

Using the assumption that $\|u\|_1 \geq \|v\|_1 - \Delta$, the triangle inequality, and the fact that $\|u_{S^c}\|_1 = \sigma_k(u)$, we obtain

$$\begin{aligned} \|u\|_1 &\geq \|v\|_1 - \Delta \\ &= \|u-y\|_1 - \Delta \\ &= \|(u-y)_S\|_1 + \|(u-y)_{S^c}\|_1 - \Delta \\ &\geq \|u_S\|_1 - \|y_S\|_1 + \|u_{S^c}\|_1 - \|y_{S^c}\|_1 - \Delta \\ &= \|u\|_1 - 2\|u_{S^c}\|_1 + \|y\|_1 - 2\|y_S\|_1 - \Delta \\ &= \|u\|_1 - 2\sigma_k(u) + \|y\|_1 - 2\|y_S\|_1 - \Delta, \end{aligned}$$

which yields

$$\|y\|_1 \leq 2\sigma_k(u) + 2\|y_S\|_1 + \Delta.$$

Using (24) to bound $\|y_S\|_1$, we further obtain

$$\|y\|_1 \leq 2\sigma_k(u) + \frac{2\|Fu - Fv\|_1 + 4d\epsilon\|y\|_1}{(1-2\epsilon)d} + \Delta.$$

Rearranging this inequality completes the proof.

Appendix C
Technical lemmas

Lemma C.1. Any $\theta \in \Theta_L$ satisfies the bound

$$\|\Phi(\theta^* - \theta)\|_1^2 \leq 4L\sum_{i=1}^{m}\left|(\Phi\theta^*)_i^{1/2} - (\Phi\theta)_i^{1/2}\right|^2.$$

Proof: From Lemma II.4 it follows that

$$\|\Phi\theta\|_1 \leq \|\theta\|_1 \leq L, \qquad \forall\theta\in\Theta_L. \tag{25}$$

Let $\beta^* \triangleq \Phi\theta^*$ and $\beta \triangleq \Phi\theta$. Then

$$\begin{aligned}
\|\beta^* - \beta\|_1^2 &= \left(\sum_{i=1}^m |{\beta^*}_i - \beta_i|\right)^2 \\
&= \left(\sum_{i=1}^m \left|{\beta^*}_i^{1/2} - \beta_i^{1/2}\right| . \left|{\beta^*}_i^{1/2} + \hat{\beta}_i^{1/2}\right|\right)^2 \\
&\le \sum_{i,j=1}^m \left|{\beta^*}_i^{1/2} - \beta_i^{1/2}\right|^2 . \left|{\beta^*}_j^{1/2} + \beta_j^{1/2}\right|^2 \\
&\le 2\sum_{i=1}^m \left|{\beta^*}_i^{1/2} - \beta_i^{1/2}\right|^2 . \sum_{j=1}^m \left|{\beta^*}_j + \beta_j\right| \\
&= 2\sum_{i=1}^m \left|{\beta^*}_i^{1/2} - \beta_i^{1/2}\right|^2 . (\|\beta^*\|_1 + \|\beta\|_1) \\
&\le 4L\sum_{i=1}^m \left|{\beta^*}_i^{1/2} - \beta_i^{1/2}\right|^2 . \\
&= 4L\sum_{i=1}^m \left|(\Phi f^*)_i^{1/2} - (\Phi f)_i^{1/2}\right|^2 .
\end{aligned}$$

The first and the second inequalities are by Cauchy–Schwarz, while the third inequality is a consequence of Eq. (25). ∎

Lemma C.2. Let $\hat{\theta}$ be a minimizer in Eq. (6). Then

$$\begin{aligned}
&\mathbb{E}_{\Phi\theta^*}\left[\sum_{i=1}^m \left|(\Phi\theta^*)_i^{1/2} - (\Phi\hat{\theta})_i^{1/2}\right|^2\right] \\
&\quad\le \min_{\theta\in\Theta_L} \left[\mathrm{KL}(\mathbb{P}_{\Phi\theta^*} \,\|\, \mathbb{P}_{\Phi\theta}) + 2\,\mathrm{pen}(\theta)\right]. \qquad (26)
\end{aligned}$$

Proof: Using Lemma C.4 below with $g = \Phi\theta^*$ and $h = \Phi\hat{\theta}$ we have

$$\begin{aligned}
&\mathbb{E}_{\Phi\theta^*}\left[\sum_{i=1}^m \left|(\Phi\theta^*)_i^{1/2} - (\Phi\hat{\theta})_i^{1/2}\right|^2\right] \\
&\quad= \mathbb{E}_{\Phi\theta^*}\left[2\log\frac{1}{\int\sqrt{\mathbb{P}_{\Phi\theta^*}(y)\mathbb{P}_{\Phi\hat{\theta}}(y)}d\nu(y)}\right].
\end{aligned}$$

Clearly

$$\int\sqrt{\mathbb{P}_{\Phi\theta^*}(y)\mathbb{P}_{\Phi\hat{\theta}}(y)}d\nu(y) = \mathbb{E}_{\Phi\theta^*}\left[\sqrt{\frac{\mathbb{P}_{\Phi\hat{\theta}}(y)}{\mathbb{P}_{\Phi\theta^*}(y)}}\right].$$

We now provide a bound for this expectation. Let $\tilde{\theta}$ be a minimizer of $\mathrm{KL}(\mathbb{P}_{\Phi\theta^*}\|\mathbb{P}_{\Phi\theta}) + 2\,\mathrm{pen}(\theta)$ over $\theta \in \Theta_L$. Then, by definition of $\hat{\theta}$, we have

$$\sqrt{\mathbb{P}_{\Phi\hat{\theta}}(y)}e^{-\mathrm{pen}(\hat{\theta})} \ge \sqrt{\mathbb{P}_{\Phi\tilde{\theta}}(y)}e^{-\mathrm{pen}(\tilde{\theta})}$$

for every y. Consequently,

$$\frac{1}{\mathbb{E}_{\Phi\theta^*}\left[\sqrt{\frac{\mathbb{P}_{\Phi\hat{\theta}}(y)}{\mathbb{P}_{\Phi\theta^*}(y)}}\right]} \le \frac{\sqrt{\mathbb{P}_{\Phi\hat{\theta}}(y)}e^{-\mathrm{pen}(\hat{\theta})}}{\sqrt{\mathbb{P}_{\Phi\tilde{\theta}}(y)}e^{-\mathrm{pen}(\tilde{\theta})}\mathbb{E}_{\Phi\theta^*}\left[\sqrt{\frac{\mathbb{P}_{\Phi\hat{\theta}}(y)}{\mathbb{P}_{\Phi\theta^*}(y)}}\right]},$$

We can split the quantity

$$2\mathbb{E}_{\Phi\theta^*}\left[\log\left(\frac{\sqrt{\mathbb{P}_{\Phi\hat{\theta}}(y)}e^{-\mathrm{pen}(\hat{\theta})}}{\sqrt{\mathbb{P}_{\Phi\tilde{\theta}}(y)}e^{-\mathrm{pen}(\tilde{\theta})}\mathbb{E}_{\Phi\theta^*}\left[\sqrt{\frac{\mathbb{P}_{\Phi\hat{\theta}}(y)}{\mathbb{P}_{\Phi\theta^*}(y)}}\right]}\right)\right]$$

into three terms:

$$\begin{aligned}
&\mathbb{E}_{\Phi\theta^*}\left[\log\left(\frac{\mathbb{P}_{\Phi\theta^*}(y)}{\mathbb{P}_{\Phi\tilde{\theta}}(y)}\right)\right] + 2\,\mathrm{pen}(\tilde{\theta}) \\
&+ 2\mathbb{E}\left[\log\left(\frac{\sqrt{\mathbb{P}_{\Phi\hat{\theta}}(y)}e^{-\mathrm{pen}(\hat{\theta})}}{\sqrt{\mathbb{P}_{\Phi\theta^*}(y)}\mathbb{E}_{\Phi\theta^*}\left[\sqrt{\frac{\mathbb{P}_{\Phi\hat{\theta}}(y)}{\mathbb{P}_{\Phi\theta^*}(y)}}\right]}\right)\right]
\end{aligned}$$

We show that the third term is always nonpositive, which completes the proof. Using Jensen's inequality,

$$\begin{aligned}
&\mathbb{E}\left[\log\left(\frac{\sqrt{\mathbb{P}_{\Phi\hat{\theta}}(y)}e^{-\mathrm{pen}(\hat{\theta})}}{\sqrt{\mathbb{P}_{\Phi\theta^*}(y)}\mathbb{E}_{\Phi\theta^*}\left[\sqrt{\frac{\mathbb{P}_{\Phi\hat{\theta}}(y)}{\mathbb{P}_{\Phi\theta^*}(y)}}\right]}\right)\right] \\
&\le \log\left(\mathbb{E}\left[\frac{\sqrt{\mathbb{P}_{\Phi\hat{\theta}}(y)}e^{-\mathrm{pen}(\hat{\theta})}}{\sqrt{\mathbb{P}_{\Phi\theta^*}(y)}\mathbb{E}_{\Phi\theta^*}\left[\sqrt{\frac{\mathbb{P}_{\Phi\hat{\theta}}(y)}{\mathbb{P}_{\Phi\theta^*}(y)}}\right]}\right]\right).
\end{aligned}$$

Now

$$\mathbb{E}\left[\frac{\sqrt{\mathbb{P}_{\Phi\hat{\theta}}(y)}e^{-\mathrm{pen}(\hat{\theta})}}{\sqrt{\mathbb{P}_{\Phi\theta^*}(y)}\mathbb{E}_{\Phi\theta^*}\left[\sqrt{\frac{\mathbb{P}_{\Phi\hat{\theta}}(y)}{\mathbb{P}_{\Phi\theta^*}(y)}}\right]}\right] \le \sum_{\theta\in\Theta_L} e^{-\mathrm{pen}(\theta)} \le 1.$$

Since $\mathbb{E}_{\Phi\theta^*}\left[\log\left(\frac{\mathbb{P}_{\Phi\theta^*}(y)}{\mathbb{P}_{\Phi\tilde{\theta}}(y)}\right)\right] = \mathrm{KL}(\mathbb{P}_{\Phi\theta^*}\|\mathbb{P}_{\Phi\tilde{\theta}})$, we obtain

$$\begin{aligned}
&\mathbb{E}_{\Phi\theta^*}\left[\sum_{i=1}^m \left|(\Phi\theta^*)_i^{1/2} - (\Phi\hat{\theta})_i^{1/2}\right|^2\right] \\
&\quad\le \mathrm{KL}(\mathbb{P}_{\Phi\theta^*}\|\mathbb{P}_{\Phi\tilde{\theta}}) + 2\,\mathrm{pen}(\tilde{\theta}) \\
&\quad= \min_{\theta\in\Theta_L} \left[\mathrm{KL}(\mathbb{P}_{\Phi\theta^*}\|\mathbb{P}_{\Phi\theta}) + 2\,\mathrm{pen}(\theta)\right],
\end{aligned}$$

which proves the lemma. ∎

Lemma C.3. If the estimators in Θ_L satisfy the condition (9), then following inequality holds:

$$\mathrm{KL}(\mathbb{P}_{\Phi\theta^*} \,\|\, \mathbb{P}_{\Phi\theta}) \le \frac{1}{c}\|\theta^* - \theta\|_1^2, \qquad \forall\theta \in \Theta_L.$$

Proof: By definition of the KL divergence,

$$\begin{aligned}
&\mathrm{KL}(\mathbb{P}_{\Phi\theta^*}\|\mathbb{P}_{\Phi\theta}) \\
&= \mathbb{E}_{\Phi\theta^*}\left[\log\left(\frac{\mathbb{P}_{\Phi\theta^*}(y)}{\mathbb{P}_{\Phi\theta}(y)}\right)\right] \\
&= \sum_{j=1}^{m}\mathbb{E}_{(\Phi\theta^*)_j}\left[y_j\log\left(\frac{(\Phi\theta^*)_j}{(\Phi\theta)_j}\right)\right] \\
&\qquad\qquad -\sum_{j=1}^{m}\mathbb{E}_{(\Phi\theta^*)_j}\left[(\Phi\theta^*)_j-(\Phi\theta)_j\right] \\
&= \sum_{j=1}^{m}\left[(\Phi\theta^*)_j\log\left(\frac{(\Phi\theta^*)_j}{(\Phi\theta)_j}\right)-(\Phi\theta^*)_j+(\Phi\theta)_j\right] \\
&\le \sum_{j=1}^{m}(\Phi\theta^*)_j\left(\frac{(\Phi\theta^*)_j}{(\Phi\theta)_j}-1\right)-(\Phi\theta^*)_j+(\Phi\theta)_j \\
&= \sum_{j=1}^{m}\frac{1}{(\Phi\theta)_j}\left|(\Phi\theta^*-\Phi\theta)_j\right|^2 \\
&\le \frac{1}{c}\|\Phi(\theta^*-\theta)\|_2^2 \\
&\le \frac{1}{c}\|\Phi(\theta^*-\theta)\|_1^2 \le \frac{1}{c}\|\theta^*-\theta\|_1^2.
\end{aligned}$$

The first inequality uses $\log t \le t-1$, the second is by (9), the third uses the fact that the ℓ_1 norm dominates the ℓ_2 norm, and the last one is by the RIP-1 property (Lemma II.4). ∎

Lemma C.4. Given two Poisson parameter vectors $g, h \in \mathbb{R}_+^m$, the following equality holds:

$$2\log\frac{1}{\int\sqrt{\mathbb{P}_g(y)\mathbb{P}_h(y)}d\mu(y)} = \sum_{j=1}^{m}\left|g_j^{1/2}-h_j^{1/2}\right|^2,$$

where μ denotes the counting measure on $\mathbb{R}_+^m$.

Proof:

$$\begin{aligned}
&\int\sqrt{\mathbb{P}_g(y)\mathbb{P}_h(y)}d\mu(y) \\
&= \prod_{j=1}^{m}\sum_{y_j=0}^{\infty}\frac{(g_jh_j)^{y_j/2}}{y_j!}e^{-(g_j+h_j)/2} \\
&= \prod_{j=1}^{m}e^{-\frac{1}{2}(g_j-2(g_jh_j)^{1/2}+h_j)}\sum_{y_j=0}^{\infty}\frac{(g_jh_j)^{y_j/2}}{y_j!}e^{-(g_jh_j)^{1/2}} \\
&= \prod_{j=1}^{m}e^{-\frac{1}{2}(g_j-2(g_jh_j)^{1/2}+h_j)}\underbrace{\int\mathbb{P}_{(g_jh_j)^{1/2}}(y_j)d\nu_j(y_j)}_{=1} \\
&= \prod_{j=1}^{m}e^{-\frac{1}{2}\left(g_j^{1/2}-h_j^{1/2}\right)^2}
\end{aligned}$$

Taking logs, we obtain the lemma. ∎

ACKNOWLEDGMENT

The authors would like to thank Piotr Indyk for his insightful comments on the performance of the expander graphs, and the anonymous referees whose constructive criticism and numerous suggestions helped improve the quality of the paper.

REFERENCES

[1] D. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289–1306, April 2006.

[2] E. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Commun. Pure Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, 2006.

[3] D. Snyder, A. Hammond, and R. White, "Image recovery from data acquired with a charge-coupled-device camera." *J. Opt. Soc. Amer. A*, vol. 10, pp. 1014–1023, 1993.

[4] D. Bertsekas and R. Gallager, *Data Networks*. Prentice-Hall, 1992.

[5] I. Rish and G. Grabarnik, "Sparse signal recovery with exponential-family noise," in *Allerton Conference on Communication, Control, and Computing*, 2009.

[6] L. Jacques, D. K. Hammond, and M. J. Fadili, "Dequantizing compressed sensing with non-gaussian constraints," in *Proc. of ICIP09*, 2009.

[7] R. E. Carrillo, K. E. Barner, and T. C. Aysal, "Robust sampling and reconstruction methods for sparse signals in the presence of impulsive noise," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 392–408, 2010.

[8] J. N. Laska, M. A. Davenport, and R. G. Baraniuk, "Exact signal recovery from sparsely corrupted measurements through the pursuit of justice," in *43rd Asilomar Conference on Signals, Systems and Computers*, 2009.

[9] R. Willett and M. Raginsky, "Performance bounds on compressed sensing with Poisson noise," in *Proc. IEEE Int. Symp. on Inform. Theory*, Seoul, Korea, Jun/Jul 2009, pp. 174–178.

[10] M. Raginsky, Z. Harmany, R. Marcia, and R. Willett, "Compressed sensing performance bounds under Poisson noise," *IEEE Trans. Signal Process.*, vol. 58, pp. 3990–4002, August 2010.

[11] E. Candès and T. Tao, "Near optimal signal recovery from random projections: Universal encoding strategies," *IEEE Trans. Inform. Theory*, vol. 52, no. 12, pp. 5406–5425, December 2006.

[12] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[13] S. Hoory, N. Linial, and A. Wigderson, "Expander Graphs and their Applications," *Bull. Amer. Math. Soc. (New Series)*, vol. 43, 2006.

[14] R. Berinde and P. Indyk, "Sparse recovery using sparse random matrices," *Technical Report, MIT*, 2008.

[15] S. Jafarpour, W. Xu, B. Hassibi, and R. Calderbank, "Efficient and robust compressed sensing using optimized expander graphs," *IEEE Trans. Inform. Theory*, vol. 55, no. 9, pp. 4299–4308, September 2009.

[16] R. Berinde, P. Indyk, and M. Ruzic, "Practical near-optimal sparse recovery in the ℓ_1 norm," *46th Annual Allerton Conf. on Comm., Control, and Computing*, 2008.

[17] J. Tropp and A. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inform. Theory*, vol. 53, no. 12, pp. 4655–4666, December 2007.

[18] R. Berinde, A. Gilbert, P. Indyk, H. Karloff, and M. Strauss, "Combining geometry and combinatorics: a unified approach to sparse signal recovery." *46th Annual Allerton Conference on Communication, Control, and Computing*, pp. 798–805, September 2008.

[19] P. Indyk and M. Ruzic, "Near-optimal sparse recovery in the $\ell 1$ norm," *Proc. 49th Ann. IEEE Symp. on Foundations of Computer Science (FOCS)*, pp. 199–207, 2008.

[20] N. Alon and J. Spencer, *The Probabilistic Method*. Wiley-Interscience, 2000.

[21] M. Sipser and D. Spielman, "Expander Codes," *IEEE Trans. Inform. Theory*, vol. 42, no. 6, pp. 1710–1722, 1996.

[22] V. Guruswami, J. Lee, and A. Razborov, "Almost euclidean subspaces of ℓ_1 via expander codes," *Proceedings of the 19th annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 353–362, January 2008.

[23] V. Guruswami, C. Umans, and S. Vadhan., "Unbalanced expanders and randomness extractors from Parvaresh-Vardy codes." *IEEE Conference on Computational Complexity (CCC)*, 2007.

[24] F. Parvaresh and A. Vardy, "Correcting errors beyond the Guruswami-Sudan radius in polynomial time," *Proc. 46th Ann. IEEE Symp. on Foundations of Computer Science (FOCS)*, pp. 285–294, 2005.

[25] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Wiley, 2006.

[26] P. McCullagh and J. Nelder, *Generalized Linear Models*, 2nd ed. London: Chapman and Hall, 1989.

[27] L. Wasserman, *All of Statistics: A Concise Course in Statistical Inference*. Springer, 2003.

[28] Z. Harmany, R. Marcia, and R. Willett, "Sparse Poisson intensity reconstruction algorithms," in *Proc. IEEE Stat. Sig. Proc. Workshop*, 2009.
[29] C. Estan and G. Varghese, "New directions in traffic measurement and accounting: focusing on the elephants, ignoring the mice," *ACM Trans. Computer Sys.*, vol. 21, no. 3, pp. 270–313, 2003.
[30] Y. Lu, A. Montanari, B. Prabhakar, S. Dharmapurikar, and A. Kabbani, "Counter Braids: a novel counter architecture for per-flow measurement," in *Proc. ACM SIGMETRICS*, 2008.
[31] W. Fang and L. Peterson, "Inter-AS traffic patterns and their implications," in *Proc. IEEE GLOBECOM*, 1999.
[32] A. Feldmann, A. Greenberg, C. Lund, N. Reingold, J. Rexford, and F. True, "Deriving traffic demands for operational IP networks: methodology and experience," in *Proc. ACM SIGCOMM*, 2000.
[33] R. Berinde, A. Gilbert, P. Indyk, H. Karloff, and M. Strauss, "Combining geometry and combinatorics: a unified approach to sparse signal recovery," in *Proc. Allerton Conf.*, September 2008, pp. 798–805.
[34] M. A. Khajehnejad, A. G. Dimakis, W. Xu, and B. Hassibi, "Sparse recovery of positive signals with minimal expansion," *Submitted*, 2009.
[35] E. Candès and J. Romberg, "ℓ_1-MAGIC: Recovery of Sparse Signals via Convex Programming," *available at http://www.acm.caltech.edu/l1magic*, 2005.
[36] R. Berinde and P. Indyk, "Sequential Sparse Matching Pursuit," *Allerton*, 2009.
[37] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 1.21," http://cvxr.com/cvx, Jan. 2011.
[38] ——, "Graph implementations for nonsmooth convex programs," in *Recent Advances in Learning and Control*, ser. Lecture Notes in Control and Information Sciences, V. Blondel, S. Boyd, and H. Kimura, Eds. Springer-Verlag Limited, 2008, pp. 95–110, http://stanford.edu/∼boyd/graph_dcp.html.

Maxim Raginsky received the B.S. and M.S. degrees in 2000 and the Ph.D. degree in 2002 from Northwestern University, Evanston, IL, all in electrical engineering. From 2002 to 2004 he was a Postdoctoral Researcher at the Center for Photonic Communication and Computing at Northwestern University, where he pursued work on quantum cryptography and quantum communication and information theory. From 2004 to 2007 he was a Beckman Foundation Postdoctoral Fellow at the University of Illinois in Urbana-Champaign, where he carried out research on information theory, statistical learning and computational neuroscience. Since September 2007 he has been with Duke University, where he is now Assistant Research Professor of Electrical and Computer Engineering. His interests include statistical signal processing, information theory, statistical learning and nonparametric estimation. He is particularly interested in problems that combine the communication, signal processing and machine learning components in a novel and nontrivial way, as well as in the theory and practice of robust statistical inference with limited information.

Sina Jafarpour is a Ph.D. candidate in the Computer Science Department of Princeton University, co-advised by Prof. Robert Calderbank and Prof. Robert Schapire. He received his B.Sc. in Computer Engineering from Sharif University of Technology in 2007. His main research interests include compressed sensing and applications of machine learning in image processing, multimedia, and information retrieval. Mr Jafarpour has been a member of the Van Gogh project supervised by Prof. Ingrid Daubechies since Fall 2008.

Zachary Harmany received the B.S. (magna cum lade) in Electrical Engineering and B.S. (cum lade) in Physics in 2006 from The Pennsylvania State University. Currently, he is a graduate student in the department of Electrical and Computer Engineering at Duke University. In 2010 he was a visiting researcher at The University of California, Merced. His research interests include nonlinear optimization, statistical signal processing, learning theory, and image processing with applications in functional neuroimaging, medical imaging, astronomy, and night vision. He is a student member of the IEEE as well as a member of SIAM and SPIE.

Roummel Marcia received his B.A. in Mathematics from Columbia University in 1995 and his Ph.D. in Mathematics from the University of California, San Diego in 2002. He was a Computation and Informatics in Biology and Medicine Postdoctoral Fellow in the Biochemistry Department at the University of Wisconsin-Madison and a Research Scientist in the Electrical and Computer Engineering at Duke University. He is currently an Assistant Professor of Applied Mathematics at the University of California, Merced. His research interests include nonlinear optimization, numerical linear algebra, signal and image processing, and computational biology.

Rebecca Willett is an assistant professor in the Electrical and Computer Engineering Department at Duke University. She completed her Ph.D. in Electrical and Computer Engineering at Rice University in 2005. Prof. Willett received the National Science Foundation CAREER Award in 2007, is a member of the DARPA Computer Science Study Group, and received an Air Force Office of Scientific Research Young Investigator Program award in 2010. Prof. Willett has also held visiting researcher positions at the Institute for Pure and Applied Mathematics at UCLA in 2004, the University of Wisconsin-Madison 2003-2005, the French National Institute for Research in Computer Science and Control (INRIA) in 2003, and the Applied Science Research and Development Laboratory at GE Healthcare in 2002. Her research interests include network and imaging science with applications in medical imaging, wireless sensor networks, astronomy, and social networks. Additional information, including publications and software, are available online at http://www.ee.duke.edu/∼willett/.

Robert Calderbank received the B.Sc. degree in 1975 from Warwick University, England, the M.Sc. degree in 1976 from Oxford University, England, and the Ph.D. degree in 1980 from the California Institute of Technology, all in mathematics.

Dr. Calderbank is Dean of Natural Sciences at Duke University. He was previously Professor of Electrical Engineering and Mathematics at Princeton University where he directed the Program in Applied and Computational Mathematics. Prior to joining Princeton in 2004, he was Vice President for Research at AT&T, responsible for directing the first industrial research lab in the world where the primary focus is data at scale. At the start of his career at Bell Labs, innovations by Dr. Calderbank were incorporated in a progression of voiceband modem standards that moved communications practice close to the Shannon limit. Together with Peter Shor and colleagues at AT&T Labs he showed that good quantum error correcting codes exist and developed the group theoretic framework for quantum error correction. He is a co-inventor of space-time codes for wireless communication, where correlation of signals across different transmit antennas is the key to reliable transmission.

Dr. Calderbank served as Editor in Chief of the IEEE Transactions on Information Theory from 1995 to 1998, and as Associate Editor for Coding Techniques from 1986 to 1989. He was a member of the Board of Governors of the IEEE Information Theory Society from 1991 to 1996 and from 2006 to 2008. Dr. Calderbank was honored by the IEEE Information Theory Prize Paper Award in 1995 for his work on the Z_4 linearity of Kerdock and Preparata Codes (joint with A.R. Hammons Jr., P.V. Kumar, N.J.A. Sloane, and P. Sole), and again in 1999 for the invention of space-time codes (joint with V. Tarokh and N. Seshadri). He received the 2006 IEEE Donald G. Fink Prize Paper Award and the IEEE Millennium Medal, and was elected to the U.S. National Academy of Engineering in 2005.